Empirical, Automated Vocabulary Discovery Using Large Text Corpora and Advanced Natural Language Processing Tools

William R. Hersh, M.D.*, Emily H. Campbell, R.N., M.S.*,
David A. Evans, Ph.D.†, Nicholas D. Brownlow, M.S.†

*Biomedical Information Communication Center
Oregon Health Sciences University
Portland, Oregon, USA

†Claritech Corp.
Pittsburgh, Pennsylvania, USA

A major impediment to the full benefit of electronic medical records is the lack of a comprehensive clinical vocabulary. Most existing vocabularies do not allow the full expressiveness of clinical diagnoses and findings that are often qualified by modifiers relating to severity, acuity, and temporal factors. One reason for the lack of expressivity is the inability of traditional manual construction techniques to identify the diversity of language used by clinicians. This study used advanced natural language processing tools to identify terminology in a clinical findings domain, compare its coverage with the UMLS Metathesaurus, and quantify the effort required to discover the additional terminology. It was found that substantial amounts of phrases and individual modifiers were not present in the UMLS Metathesaurus and that modest effort in human time and computer processing were needed to obtain the larger quantity of terms.

INTRODUCTION

One of the major promises of the electronic medical record (EMR) is the ability to utilize patient data on individual and aggregate levels to enhance care. On the individual level, data can be used for decision support, such as reminding clinicians to perform preventive measures [1] or alert them to dangerous situations [2], while on the aggregate level, it can be used to enhance research and quality assurance [3]. A major problem with current EMR systems, however, is the lack of clinical vocabulary that can express the information in patient diagnoses and findings [4, 5]. Diagnoses, for example, are coded in ICD-9-CM, which cannot capture aspects such as the severity of an illness (e.g., whether a cancer is localized or metastatically spreading) or its chronicity [6]. Clinical findings in the history and physical exam are even more problematic, due to the large number of attributes that modify each finding [7]. Chest pain, for example, takes on a different meaning to clinicians when it is crushing and comes with exertion as opposed to when it is burning and comes when lying down.

Thus until more detailed methods for representing information about the patient can be developed, the true potential of the EMR will remain untapped. One of us has outlined the requirements for clinical vocabularies [7], which include:

1. Lexical decomposition - to allow the meaning of individual words to be discerned in the context of the entire term.
2. Semantic typing - to allow for identification of synonyms and translation across semantic "equivalence classes."
3. Compositional extensibility - to allow words to be combined to generate new concepts.

Most existing vocabularies do not meet these criteria, as they consist of term lists, usually created by consensus panels of experts in given fields. The SNOMED vocabulary comes closest in allowing terms from different axes to be combined [8], but there is no constraining grammar to insure legitimate terms arise from combinations of words from different axes nor any means to prevent redundancy [4].

A major impediment to vocabularies based on the above requirements is the complexity of building and maintaining a lexicon and grammar for medical concepts and their constituent parts, especially one which represents the differing conceptual views of multiple specialties within the health care field [9]. Our view is that the most serious problem has been the lack of tools to capture the quantity and diversity of terminology used by clinicians that would serve as the basis for such vocabularies. The major goal of this study was to assess the ability of advanced natural language processing (NLP) tools to assist in the empirical creation of clinical vocabularies using large quantities of EMR text. The steps in this process included:

0195—4210/96/$5.00 © 1996 AMIA, Inc.

1. Assessing the degree of terminology not present in a large clinical vocabulary (the UMLS Metathesaurus).
2. Determining the nature and generalizability of words modifying noun phrase (NP) heads as determined by their semantic types and their occurrence with different heads respectively. (The head of an NP is generally the word that occurs at the end of phrase and represents the central concept. Other words in the NP are modifiers to the head, which are usually adjectives or other nouns.)
3. Measuring the effort required to obtain this additional terminology.

The advanced NLP tool was the CLARIT system (Claritech Corp., Pittsburgh, PA) [10]. Like many modern NLP systems, CLARIT foregoes the unattainable goal of unambiguous recognition of English text. Rather, it aims to identify the constituents of language most important in understanding the conceptual content of text, namely NPs. CLARIT can parse text at a rate of 2-3 megabytes per minute, making it possible to extract NPs over very large quantities of text. These NPs were used to carry out the steps listed above.

METHODS

The corpus for parsing in this experiment was all narrative reports available in the EMR systems at Oregon Health Sciences University and the Portland Veterans Administration Medical Center through February, 1995, a total of 842 megabytes. This text included all dictated reports, including discharge summaries, outpatient progress notes, emergency room notes, radiology reports, and letters. The vocabulary used for the coverage comparison studies was the 1996 UMLS Metathesaurus.

For the analysis in this paper, we aimed to focus on a broad area of language, oriented to clinical findings, that was frequently used by health care practitioners. We consulted a book on the topic of common patient complaints [11] and noted that 16 of the 31 chapters dealt with some aspect of pain (e.g., abdominal pain, backache, earache, heartburn, dysuria, etc.). We developed a list of 10 words (or morphemes within words) representing NP heads dealing with pain: ache, algia, dysmenorrhea, dyspepsia, dysphagia, dysuria, pain, sore, tender, and throb. Some of the terms had prefixes (e.g., *back*ache, *my*algia) while others had suffixes (e.g., sore*s*, tender*ness*). For tabulating results, we counted NPs and words for the 10 heads and did not subclassify those with prefixes or suffixes.

For assessing the degree of terminology not present in the UMLS Metathesaurus, we counted the number of complete terms as well as the number of words modifying each of the 10 NP heads in the EMR corpus and UMLS Metathesaurus. As medical records have a significant number of misspellings, we eliminated all words not present in one of three clinical vocabularies (the UMLS Metathesaurus, SNOMED, and the Medical Entities Dictionary [6]) or the Unix spell-checker from the analysis.

To determine the nature and generalizability of modifiers, we assigned semantic types for each word in order to determine what types of words tended to occur among different types of NP heads. We began with a list of semantic types developed for chest x-ray findings [12], which needed some additions to account for terms discovered in this analysis. This provided an analysis of the types of modifiers that comprise clinical findings. To assess generalizability of these modifiers, we determined the degree to which individual modifiers occurred with more than one NP head. This was done by counting how many of the 10 NP heads each word modified.

For measuring the effort required to obtain this added terminology, we monitored the time taken to implement various utilities (e.g., the several Perl programs written to extract and count words and terms) and perform manual tasks (e.g., assigning semantic types).

RESULTS

The parsing of the EMR corpus yielded over 30 million NPs. A total of 35,316 NPs contained one of the 10 designated heads. There were 3,741 unique words in these NPs, with 491 words designated as misspellings and removed from further analysis. A total of 1,038 NPs in the UMLS Metathesaurus contained one of the 10 heads, with 375 unique words occurring as modifiers. The number of NPs and modifiers for each individual NP head are shown in Table 1. That there were a great deal more NPs in the EMR corpus than the Metathesaurus was not surprising, since clinicians are likely to use a wider diversity of phrases than those found in analysis also revealed that much larger numbers of modifiers were used for these NP heads in clinical charts as well, indicating that the breadth of modifier coverage to represent the types of findings reported by clinicians was not present in the Metathesaurus.

Analysis of the words in preparation for semantic typing revealed that three additional semantic types were required to be added to Friedman's original classification for chest x-ray findings [12]. We also subdivided her Bodyloc type into a Bodypart, which listed specific body parts, and Bodyregion, which specified regions in adjectival form. Table 2 lists the semantic types used for this analysis.

The generalizability of modifiers is presented in Table 3, which lists the number of words from each semantic category that occured with one to ten of the NP heads. While the majority of modifiers were limited to one head, it can be seen that a substantial number occurred in multiple terms. The bottom row lists the total

most common types were descriptors (e.g sharp, stabbing) and body regions (e.g cardiac, cervical), both of which are likely t be generalizable across domains. Table 4 list the most frequent modifiers, i.e., those th occurred with 8, 9, or 10 heads.

The time required for this analysis was mostl human effort devoted to programming an assigning semantic types. All of the utilitie used to extract and manipulate data for thi experiment were implemented by a singl programmer using the Perl programmin language in a period of two months time Once programmed, the analyses ran quickly CLARIT parsed all the NPs in the corpus i under 14 hours on a 75 Mhz Sun Sparcstatio 20 with 128 MB of RAM. The Perl utilitie ran through the remaining automated analyse in under 12 hours of machine time. Th semantic typing required about 50 person hours.

Table 1 -- Noun phrases and modifiers for each NP head in the EMR corpus and UMLS Metathesaurus.

NP Head	Corpus NPs	Corpus Modifiers	Metathes. NPs	Metathes. Modifiers
ache	2,239	804	47	38
algia	421	226	115	47
dysmenorrhea	32	35	7	5
dyspepsia	89	76	3	1
dysphagia	175	165	10	9
dysuria	112	101	4	3
pain	11,554	2,548	290	270
sore	514	328	20	12
tender	3,704	1,026	87	94
throb	41	38	1	1

Table 2 -- Semantic types, based on Friedman's classification for chest x-ray findings [12], with abbreviations denoted for Table 3. (* added to original classification)

Bodypart - Terms that specify a part of the body. (BP)*
Bodyregion - Terms that specify a well-defined area of the body. (BR)*
Certainty - Terms affecting the certainty of a finding. (CTY)
Cfinding - Terms denoting a complete finding because these terms implicitly or explicitly contain a finding and a body location.
Change - Terms denoting a change in findings where the change is an improvement or worsening of a finding but not the start or end. (CHG)
Chemical/Drug - Terms denoting a chemical or drug. (CHEM)*
Connector - Terms that connect one finding to another. (CONN)
Degree - Terms denoting the severity of a finding.
Descriptor - Terms qualifying a property of a body location or finding. (DESC)
Device - Terms denoting devices. (DEV)
Disease - Terms denoting a disease. (DIS)
Otherpartofspeech - Articles, determiners, etc..*
Position - Terms denoting orientation. (POS)
Pfinding - Terms denoting a partial finding.
Procedure - Terms denoting a therapeutic or diagnostic procedure. (PROC)
Quantity - Terms representing non-numeric quantitative information. (QTY)
Recommend - Terms denoting recommendations.
Region - Terms denoting relative locations within a body location.
Status - Terms denoting temporal information other than an improvement or worsening of a finding. (STAT)
Technique - Terms denoting information related to the manner in which a finding was obtained.
Verb - Verbs not appearing in adjectified form.*

Table 3 -- Number of modifiers occurring in one to ten NP heads by semantic type. The ALL column sums the total of all columns to its left, while the OTHER column lists the totals for all other semantic types.

Heads	BP	BR	CTY	CHG	CHEM	CONN	DESC	DEV	DIS	POS	PROC	QTY	STAT	ALL	OTHER
1	10	248	11	10	37	6	1036	7	115	48	26	10	5	1569	518
2	10	136	8	6	6	6	301	2	18	18	3	9	0	523	147
3	2	42	6	4	0	2	119	1	5	7	0	1	1	190	41
4	0	12	0	3	0	1	47	0	0	7	0	4	0	74	27
5	2	4	2	1	0	0	32	0	1	3	0	0	0	45	13
6	0	5	2	0	0	0	12	0	0	3	0	2	1	25	4
7	0	0	0	1	0	1	8	0	0	2	0	0	0	12	2
8	0	0	1	1	0	0	12	0	0	0	0	1	0	15	2
9	0	0	0	1	0	1	4	0	0	0	0	0	0	6	0
10	0	0	0	1	0	0	1	0	0	0	0	0	0	2	0
Total	24	447	30	28	43	17	1572	10	139	88	29	27	7	2461	765

Table 4 -- The modifiers that occured in the highest number of pain NP heads.

10 heads -- increased, severe

9 heads -- chronic, increasing, mild, persistent, significant

8 heads -- associated, current, intermittent, marked, minimal, moderate, negative, occasional, ongoing, possible, progressive, recurrent, worsening

DISCUSSION

We have demonstrated that advanced NLP tools can be used to assist in the empirical discovery of new terminology from large text corpora in a predominantly automated fashion. This study has shown that this approach can identify a large quantity of terms not in the UMLS Metathesaurus as well as modifiers generalizable across multiple concepts. This process also works efficiently, indicating that porting to other domains will be feasible. The most time-consuming process was semantic typing.

The next step in this work will be to enhance real-world vocabularies with these tools. As this work is part of the National Library of Medicine's Applied Research on the Electronic Medical Record initiative, we are collaborating with other investigators who are manually developing and enhancing vocabularies. The most likely long-term role for these tools will be to provide data for the intellectual side of vocabulary construction that cannot be automated, namely creation of hierarchical and synonym classifications.

Another assessment of the vocabularies created by these tools will be to assess their use in application domains. This will be done in a series of projects assessing the ability of CLARIT and other tools to extract information from clinical narratives, including asthma progress notes to identify findings for input to practice guidelines, upper GI endoscopy reports to identify findings that predict Barrett's Esophagus, and discharge summaries to assess appropriateness of blood transfusion.

Acknowledgments

This study was supported by award U01-LM05879 from the National Library of Medicine.

References

1. McDonald CJ, Hui SL, Smith DM. Reminders to physicians from an introspective computer medical record. *Ann Int Med* 1984;100:130-138.

2. Bates DW, Kuperman G, Teich JM. Computerized physician order entry and quality of care. *Quality Management in Health Care* 1994;2:18-27.

3. Dick RB, Steen EB, eds. *The Computer-Based Patient Record: An Essential Technology for Patient Care*. Washington, DC: National Academy Press, 1991.

4. Evans DA, Cimino JJ, Hersh WR, Huff SM, Bell DA. Toward a medical concept representation language. *J Am Med Informatics Assoc* 1994;1:207-217.

5. Friedman C, Huff SM, Hersh WR, Pattison-Gordon E, Cimino JJ. The Canon Group's effort: working towards a merged model. *J Am Med Informatics Assoc* 1995;2:4-18.

6. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J Am Med Informatics Assoc* 1994;1:35-50.

7. Evans DA, Rothwell DJ, Monarch IA, Lefferts RG, Cote RA. Toward representation for medical concepts. *Med Dec Making* 1991;11:S102-S107.

8. Cote RA, Rothwell DJ, Beckett RS, Palotay JL, eds. *SNOMED International - Introduction*. Northfield, IL: College of American Pathologists, 1993.

9. Tuttle MS. The position of the Canon Group: a reality check. *J Am Med Informatics Assoc* 1994;1:298-299.

10. Evans DA, Hersh WR, Monarch IA, Lefferts RG, Handerson SK. Automatic indexing of abstracts via natural language processing using a simple thesaurus. *Med Dec Making* 1991;11:S108-S115.

11. Seller RH. *Differential Diagnosis of Common Complaints*. (Second Edition ed.) Philadelphia: W.B. Saunders, 1993.

12. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Informatics Assoc* 1994;1:161-174.